Pràctica2 - M2.951\_20221

Gerard Ramos Gambús: [gambusrg@uoc.edu](mailto:gambusrg@uoc.edu)

Oriol Caravaca Müller: ocaravacam@uoc.edu

03/01/2023

Contingut

[Descripció del *dataset* 2](#_Toc124083449)

[Integració i selecció 4](#_Toc124083450)

[Neteja de les dades 4](#_Toc124083451)

[Anàlisi de les dades 4](#_Toc124083452)

[Representació dels resultats 4](#_Toc124083453)

[Resolució del problema 4](#_Toc124083454)

[Codi 4](#_Toc124083455)

[Video 4](#_Toc124083456)

# Descripció del *dataset*

El *dataset* triat és “[Heart Attack Analysis & Prediction dataset](https://www.kaggle.com/datasets/rashikrahmanpritom/heart-attack-analysis-prediction-dataset)”, publicat per Rashik Rahman a [Kaggle](https://www.kaggle.com/datasets/rashikrahmanpritom/heart-attack-analysis-prediction-dataset) amb llicencia CC0 1.0 Universal (CC0 1.0) *Public Domain Dedication.* Aquest *dataset* conte 13 variables corresponents a dades demogràfiques i resultats de proves mèdiques cardiovasculars de 303 pacients. Així mateix, ens ofereix una variable objectiu que ens indica quin es el risc de patir un infart de cada un d’aquest pacients.

Aquest *dataset* és d’interès perquè ens permet detectar factors de risc i crear models predictius que poden ajudar a la diagnosis i detecció de problemes cardiovasculars.

A continuació es detalla el tipus i descripció de les variables del *dataset*:

|  |  |  |
| --- | --- | --- |
| Variable | Tipus | Descripció |
| age | Categòrica | Edat de la persona |
| sex | Categòrica | Gènere de la persona |
| cp | Categòrica | Tipus de dolor toràcic  1= angina típica 2 = angina atípica 3 = dolor no anginós 4 = asimptomàtic |
| trtbps | Numèrica | Pressió arterial en repòs (en mm Hg) |
| chol | Numèrica | Colestorol en mg/dl obtingut mitjançant el sensor IMC |
| fbs | Categòrica | Sucre en sang en dejú > 120 mg/dl (1 = cert; 0 = fals) |
| restecg | Categòrica | Resultats electrocardiogràfics en repòs  0 = normal 1 = tenir una anomalia de l'ona ST-T 2 = mostra una hipertròfia ventricular esquerre probable o definitiva |
| thalachh | Numèrica | Freqüència cardíaca màxima aconseguida |
| exng | Categòrica | Angina induïda per l'exercici (1 = sí; 0 = no) |
| oldpeak | Numèrica | Depressió del ST induïda per l'exercici en relació amb el repòs |
| slp | Categòrica | El pendent del segment ST de l'exercici màxim 0 = sense pendent 1 = pla 2 = pendent avall |
| caa | Numèrica | Nombre de arteries principals (0-3) |
| thall | Categòrica | Talassèmia 0 = nul 1 = defecte fixat 2 = normal 3 = defecte reversible |
| output | Categòrica | 0 = menys probabilitat d'atac cardíac  1 = més probabilitat d'atac cardíac |

# Integració i selecció

Per a la realització d’aquesta practica es contempla la utilització del *dataset* complet amb l’objectiu de fer-ne un anàlisis i implementar un model de regressió logística.

De l’estudi del model de regressió logística se’n descarta la variable fbs per considerar-se insignificant amb un nivell de significació per sota del 95%.

# Neteja de les dades

El que s’ha realitzat per a la neteja de les dades ha estat fer comprovacions sobre tot el conjunt de dades, tenint en compte si hi havia espais en blanc en les observacions o valors NA. El que s’ha vist és que les dades son consistents i no tenen valors perduts.

D’altra banda sí que podem afirmar que el conjunt de dades conté valors atípics.

De les variables que més valors atípics contenien, el que s’ha fet és identificar-los i tractar-los utilitzant la funció kNN de R, la qual utilitza l’algorisme K-nearest neighbours per reomplir amb valors que tinguin sentit les observacions que son outliers.

# Anàlisi de les dades

En primer lloc el que s’ha fet és fer un anàlisi visual per comprendre com son i com estan relacionades les dades entre si.

# Representació dels resultats

# Resolució del problema

# Codi

# Video

|  |  |
| --- | --- |
| **Contribucions** | **Signatura** |
| Investigació prèvia |  |
| Redacció de les respostes |  |
| Desenvolupament del codi |  |
| Participació al vídeo |  |